Odds of Successful Transfer of Low-level Concepts: A Key Metric for Bidirectional Speech-to-speech Machine Translation in DARPA's TRANSTAC Program

Gregory A. Sanders[1], Sébastien Bronsart[1], Sherri Condon[2], Craig Schlenoff[1]

National Institute of Standards and Technology[1]
Gaithersburg, MD 20899–8940

The MITRE Corporation[2]
McLean, VA 22102

E-mail: gregory.sanders@nist.gov, sebastien.bronsart@nist.gov, scondon@mitre.org, craig.schlenoff@nist.gov

Abstract

The Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program is a Defense Advanced Research Agency (DARPA) program to create bidirectional speech-to-speech machine translation (MT) that will allow U.S. Soldiers and Marines, speaking only English, to communicate, in tactical situations, with civilian populations who speak only other languages (for example, Iraqi Arabic). A key metric for the program is the odds of successfully transferring low-level concepts, defined as the source-language content words. The National Institute of Standards and Technology (NIST) has now carried out two large-scale evaluations of TRANSTAC systems, using that metric. In this paper we discuss the merits of that metric. It has proven to be quite informative. We describe exactly how we defined this metric and how we obtained values for it from panels of bilingual judges—allowing others to do what we have done. We compare results on this metric to results on Likert-type judgments of semantic adequacy, from the same panels of bilingual judges, as well as to a suite of typical automated MT metrics (BLEU, TER, METEOR).

1. Overview

The Spoken Language Communication and Translation System for Tactical Use (TRANSTAC) program is a Defense Advanced Research Projects Agency (DARPA) advanced technology research and development program. The goal of the TRANSTAC program is to demonstrate capabilities to rapidly develop and field free-form, two-way speech-to-speech machine translation systems that enable speakers of different languages to communicate with one another without a human interpreter — particularly to allow English-speaking U.S. Soldiers and Marines in real-world tactical situations to communicate with civilian populations who speak other languages. While the military personnel will be trained to use the systems, the assumption is that the foreign language users will have little or no chance to become familiar with using the system in advance. To date, several prototype TRANSTAC systems have been developed and formally evaluated (Weiss, et al., 2008) for force protection, civil affairs, and medical screening domains in Iraqi Arabic, and in an Indo-European "surprise language" for which the system developers were given 90 days to create a system for evaluation. Systems have been demonstrated on both PDA and laptop-grade platforms with varying performance.

A key metric for the TRANSTAC program has been the odds of successful transfer of low-level concepts (elements of meaning) from the source-language spoken input, into the target-language output, via automatic speech recognition (ASR) pipelined with machine translation (MT). The National Institute of Standards and Technology leads the evaluation team for TRANSTAC. We were asked to define a reasonable measure of the transfer of low-level concepts, and we have chosen to consider the low-level concepts to be the source-language content words or open-class words (nouns, verbs, adjectives, adverbs) plus important quantifiers and prepositions. Transfer of the low-level concepts is scored one concept at a time, as successfully transferred, deleted, or substituted (judges can also identify inserted concepts). NIST has now carried out two large-scale evaluations of TRANSTAC systems including that metric. In this paper we discuss the merits of the metric. It has proven to be quite informative.

In this paper, we take a look back at exactly how we defined the metric and at how we obtained values for it from panels of bilingual human judges — which should enable others to do what we did. To examine the performance of this metric, we present comparisons of the results from the metric to the results from other MT metrics on the same data, comparing this metric to Likert-type judgments of semantic adequacy from the same panels of bilingual human judges as well as to a suite of typical automated MT metrics (BLEU, TER, METEOR).

DARPA wants the overall evaluation to predict the end goal of the TRANSTAC program: the deployed use of speech- to-speech Machine Translation technology that enables consistently successful communication between U.S. military users and local foreign civilians whom they encounter. The TRANSTAC community is in agreement that measures should focus on (1) the semantic adequacy of the translations, leading to justified user confidence in the system's translations, and (2) the ability of an English speaker and foreign language speaker to successfully carry out a task-oriented dialogue in a narrowly focused domain of known operational need under conditions that reasonably simulate use in the field. That's the scientific problem that we are attempting to solve. This metric is a piece of that attempt

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2008	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Odds of Successful Transfer of Low-level Concepts: A Key Metric for Bidirectional Speech-to-speech Machine Translation in DARPAs TRANSTAC Program	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **National Institute of Standards and Technology (NIST) Gaithersburg, MD 20899 8940**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**7**	

2. Defining the Metric

For purposes of these metrics, pre-recorded digital audio recordings of the source language utterances were input to the MT systems being evaluated, and the resulting output of target-language text that the systems would feed to their text-to-speech (TTS) module was captured. Careful transcriptions of the source-language utterances (as well as target-language professional human translations) were also obtained. We then defined our MT metrics over textual data: the transcriptions of the input source-language utterances and the output textual target-language that would be fed to the TTS.

We have tried to define low-level concept transfer metrics, with measures of both efficiency (concepts transferred per minute) and accuracy.

Regarding efficiency, by the time we ran the most recent full-scale evaluation (in July 2007), the portion of the evaluation that used pre-recorded audio inputs was being run with no TTS output being played out. Thus, the systems could inhale the pre-recorded digital audio recordings as rapidly as they were able to process the data: they were free to accept the input more rapidly than any human could speak it, and did not have to wait for their TTS module to speak the outputs. It turned out that, for a Windows laptop platform, even the slowest system was able to process the data about as rapidly as any human would be able to speak the source-language inputs, let alone the time that would be required for the system to speak out the target language outputs. So, the transfer rate metric is uninteresting and not discussed further here.

We have stated the accuracy metric as *odds* of successful transfer of a low-level concept, which we define as the number of concepts successfully transferred divided by the number of errors. In mathematical terms, for an event, *C*, with probability P(*C*),

Odds of C = P(C) / (1 – P(C))

For example, if a six-sided die is thrown once, the odds are 5 to 1 (or 5.0) that the top face of the die will be something other than a six. Computing odds allows us to compare performance from one eval to the next as an "odds ratio", which is a fairly widely understood statistic. As P(*C*) approaches 1.0, a small increase in P(*C*) will produce a large increase in odds (for example, if P(*X*) increases from 0.98 to 0.99 then the odds will increase from 49.0 to 99.0), and that behavior causes difficulties for comparisons with other metrics that behave more like P(*X*) than like odds of *X*. Further, because we count insertions as errors, our odds computation is not quite P(*correct*) / (1 – P(*correct*)). To address these two difficulties, for purposes of this paper we have converted the odds to an adjusted P(*correct*) using straightforward algebra:

AdjP(*corr*) = 1 – (1 / (odds + 1)).

AdjP(*corr*) is more tractable for purposes of this paper such as correlations between metrics.

As our low-level concepts, we chose to count the open-class content words (nouns, verbs, adjectives, adverbs) plus those prepositions and quantifiers that the the source-language annotator considers to convey important content. If the source-language utterance was not fluent, only the words that would be present in a fluent rendition of the utterance were considered. We developed detailed guidelines for doing this source-language analysis. In all cases, the source-language concepts were identified by a well-educated native speaker of the source-language who had some formal linguistics training.

The target language analysis was performed by a panel of several bilingual judges, each a literate, educated, native speaker of the target language, using a software tool that presented the transcription of the source-language utterance at the top of the screen, followed by the target-language MT output, with the pre-identified source-language concepts presented as a vertical list (one concept per line) in a window below. For each concept, each bilingual judge scored it as correctly transferred, deleted, or substituted (source-language "A dog is on the mat" and target language "A cat is on the mat" exemplifies a substitution error). The bilingual judges could also mark elements of the target language utterance as insertions. In practice, the bilingual judges found the software tool clear and easy to use.

3. Other Metrics We Calculated

The reader will note that our metric, the odds of successful transfer of a low-level concept, is a quantitative metric of the pipelined accuracy of ASR+MT. It counts all concepts equally. Although it measures how much of the content of the source language content is also present in the target language output, there is more to the picture. Suppose, for example, the source-language utterance is, "There are now landmines buried under the road to Fallujah", and suppose the target-language translation is "There are no landmines buried under the road to Fallujah." This will score well on the odds of successful transfer of low-level elements of meaning, but the translation is terribly wrong. A fuller picture requires this metric to be paired with some complementary metric that weighs the importance of the errors and is not purely quantitative. For us, that complementary metric is a Likert-type judgment of semantic adequacy from the same panel of bilingual judges. We asked them to provide a Likert-type judgment for each utterance immediately after they scored the transfer of the low-level elements of meaning for the utterance, choosing from a four-point scale:

Completely_adequate,
Tending_adequate,
Tending_inadequate,
Inadequate.

The most widely accepted metric for MT quality is a judgment of semantic adequacy from bilingual human judges. We treat our Likert-type score as our *benchmark* score and compare our other metrics to it.

We also calculated a suite of commonly used automated metrics, intended to enable the developers to better understand the performance of their systems. The automated metrics focus on the core technologies.

For speech recognition, we calculated Word-Error-Rate (WER) — using SCTK[1] version 2.2.2 and the standard NIST procedures for normalizing the hypothesis and reference texts, thus giving English WER values that should be directly comparable to previous large-scale NIST evaluations of automatic speech recognition. MITRE normalized the non-English texts similarly, making use of their in-house expertise in those languages. For machine translation, we calculated three commonly used automated MT metrics: BLEU (Papineni *et al.*, 2001), METEOR (Banerjee and Lavie, 2004; Lavie *et al.*, 2005), and TER (Snover *et al.*, 2005), with both reference and hypothesis texts normalized much like they were normalized for ASR (see Condon, et al., 2008). We calculated BLEU, using the IBM script that had been used for the DARPA Babylon project. We calculated Translation Edit Rate (TER) scores using TerCom, version 6b, developed by Matt Snover in collaboration with BBN and the University of Maryland. We also calculated METEOR (which we modified to normalize Arabic text to some degree). METEOR was run in the mode where it scores only exact matches (no stemming or synonymy) because we could not do stemming or synonymy comparably in Arabic and the surprise language. The current standard version of METEOR is available from the dedicated METEOR web site:

http://www.cs.cmu.edu/~alavie/METEOR/

There is, in fact, a long history of various automated metrics, as well as various human-mediated metrics for the evaluation of machine translation (for example, Frederking and Nirenburg, 1994; Knight and Chander, 1994; King 1996; Hogan and Frederking, 1998; Niessen *et al.*, 2000; Frederking, et al., 2002; Melamed, Green, and Turian, 2003; Lita, Rogati, and Lavie, 2005; Russo-Lassner, Lin, and Resnik, 2005; Pozar and Charniak, 2006).

4. Discussion of the Results

Between January 2007 and July 2007, the TRANSTAC developers made substantial performance improvements at translating from spoken English to spoken Iraqi Arabic. For example, looking at the odds of successful transfer of a low-level concept, for the English-to-Arabic direction , the median value over the five systems improved from 1.55 to 4.32 (an odds ratio of 2.79), and for the Arabic to English direction, the median value over the five systems improved from 2.46 to 3.15 (an odds ratio of 1.28).

For a panel of five or six bilingual (English and Iraqi Arabic) judges, each of whom was a native speaker of Iraqi Arabic, some judges were rather more forgiving than others when scoring the successful transfer of low-level concepts (the inter-judge variation was larger, and thus that problem is larger, for the Likert-type judgments of semantic adequacy). For our panels of judges, most performance improvements in the *odds of successful transfer of low-level concepts* were large enough to be identified confidently.

[1] Available from http://www.nist.gov/speech/tools

We intended this evaluation to test the following three hypotheses.

(1) We expected strong positive correlation between the odds of successful transfer of a low-level concept and Likert-type judgments of semantic adequacy, *on average*.

The metric for odds of successful transfer of low-level concepts turned out to be strongly correlated to the percent of utterances to which the judges assigned a Likert-type semantic adequacy score of "Completely adequate". Because the percent of utterances judged to be "Completely adequate" is a probability sort of number, we have used the AdjP(*corr*) version of the odds, as explained in section 2, above. The following table shows Pearson correlations, over the five systems, between AdjP(*corr*) and the percent of utterances judged to be completely adequate (%Adeq) as well as the percent judged completely adequate minus the percent judged inadequate (%Adeq – %Inad).

	%Adeq	%Adeq – %Inad
English-to-Iraqi Arabic	0.997	0.989
Iraqi Arabic-to-English	0.978	0.982
English-to-SurpriseLang	0.997	0.994
SurpriseLang-to-English	0.960	0.990

Table 1: AdjP(*corr*) vs. Semantic Adequacy Judgments.

(2) We expected that *some* utterances would score well on odds of successful transfer of low-level concepts, but score badly on Likert-type judgments of semantic adequacy. For the reasons explained at the beginning of section 3, above. We did not expect to see the opposite relationship occur.

Having examined the system outputs, it is not clear that we could determine whether any utterances exhibited the kind of problem described at the beginning of section 3; that is, we did not find any clear-cut examples.

(3) We *hoped* to find positive correlation between our low-level concept transfer metric and existing automated MT metrics (such as BLEU and METEOR) that are known to be useful measures of MT performance for statistical MT systems.

We did find positive correlation, but the patterns observed are more complex than expected, and the patterns raise interesting questions about just what the various metrics measure. To us, this is the most interesting aspect of our findings, because it raises interesting questions about how to interpret the results from the various automated metrics versus our low-level concept transfer metric.

In order to make the relationships among all these metrics easier to understand, they are presented graphically in Figures 1 and 2, following. We created those figures as follows. For each metric separately, for each direction separately (to/from English), and for each language pair separately (Arabic⇔English and SurpriseLang⇔English) we calculated the mean and standard deviation, and then converted each value to a standard normal distribution *z*-statistic. The result is shown in Figure 1 (for Arabic) and

Figure 2 (for the surprise language). Correlations among the metrics are shown in Tables 2 – 4.

To explain Figures 1 and 2 more fully, let us elaborate on Figure 1, which shows the results for translations to/from Iraqi Arabic. There are five systems, as labeled at the bottom of the figure. For each system there are two clusters of bars, the first cluster is for the English to Iraqi Arabic direction (E2A), and the second for Arabic to English (A2E). Each cluster begins with a bar for the Odds metric that this paper focuses on — presented in its "Adjusted Fraction Correct" conversion, followed by bars for two versions of our benchmark (human judgments of semantic adequacy), and then automated MT and ASR metrics. The Translation Error Rate (TER) metric for MT and the Word Error Rate (WER) metric for automatic speech recognition have been subtracted from 1, so that all the metrics are presented with higher values being better.

The crucial point of these figures is that if two metrics give the same information, one would expect to see them have matching *z*-scores (matching bar heights) within each cluster of bars. But that did not quite happen.

The reader will note the following two patterns. First, the bars in each cluster tend to have similar heights, although the bars for TER and WER do not correlate as closely as do the others. Second, the bars for the metric that is the topic of this paper (Adjusted Fraction Correct) show a fairly strong correlation to the bars for BLEU and METEOR and to those for our benchmark human judgments. That correlation, however, is not the whole story, as can be seen by looking at the result for Sys_M translating from Iraqi Arabic to English, where the bars for our benchmark human judgments of semantic adequacy are higher than all other bars in that cluster.

We believe the explanation for that somewhat surprising datapoint may possibly be that it reflects more insertions of low-level concepts, which are reflected in the metrics, but which our bilingual human judgments of semantic adequacy may have considered unintelligible stuff that the hearer would ignore. We cannot, however, be sure what this surprising "semantic adequacy" datapoint means.

5. Suggestions on Repeatability

Our data for low-level concept transfer and for judgments of semantic adequacy represent human judgments. There can be considerable disagreements between human judgments, so that it is useful to have several judges.

It became clear even in preliminary proof-of-concept experiments that there would be disagreements among the four possible human judgments for semantic adequacy, particularly regarding Tending_inadequate vs. Inadequate. In order to improve the level of agreement, we assembled a reasonable set of utterance translations to and from Arabic, exemplifying each of the four possible judgments. This set of exemplars was drawn from utterance translations in the January 2007 evaluation on which the judges had a high level of agreement. We used that set of exemplars to train the July 2007 judges. Despite this, for the July 2007 evaluation Cohen's kappa (Cohen, 1960) for pairs of Arabic judges on the four possible semantic adequacy judgments ranged from 0.178 to 0.435 (median 0.294), which is quite low. If, however, one relaxes the criteria for a match so as to include disagreements by only one category (for example, Completely_adequate vs. Tend_adequate), the picture looks better: the range of kappa coefficients then ranges from 0.508 to 0.805 (with median 0.611), which we regard as an acceptable level of agreement. Considering all this, we suggest that a reasonably large set of judges is necessary, as outlier judges could occur. In the July 2007 evaluation we had six bilingual judges for the Arabic data and five different bilingual judges for the SurpriseLanguage data.

For the judgments of semantic adequacy, we counted occurrences of each of the four possible judgments for each system. This combination of training the judges with a set of exemplars plus averaging over several judges seems to us to be a reasonable way of dealing with less than complete inter-judge agreement.

With respect to the low-level concept transfer judgments (from which we calculated the odds of successful transfer of a low-level concept), having five or six judges gave us a fairly close level of agreement between the mean and median values for odds over the judges. Trimmed means are also a possibility if one has at least five judges.

6. Future Work

Work is underway on generating HTER (Snover, et al., 2006) values for the Arabic dataset, but this has not yet happened. HTER counts edits by a human posteditor.

Collaborative work is underway at NIST and MITRE to enhance METEOR for Arabic with morphological analysis, which we believe will help. A port of WordNet to Arabic is underway at MITRE and may possibly be incorporated into our version of METEOR.

7. Conclusions

We have presented our metric for the odds of successful transfer of a low-level concept, explaining how it was defined and how the values were generated. We have presented results for the metric in three widely disparate languages, and a variety of subject domains. Results presented included correlations with common automated MT metrics and with human judgments of semantic adequacy. We have discussed the basically quantitative nature of this metric (a characteristic shared by the automated MT metrics, and for that matter by HTER) and pointed out that human judgments of semantic adequacy do not share this purely quantitative characteristic.

8. Disclaimer

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology (NIST), nor is it intended to imply that the equipment, instruments, software, or materials are necessarily the best available for the purpose.

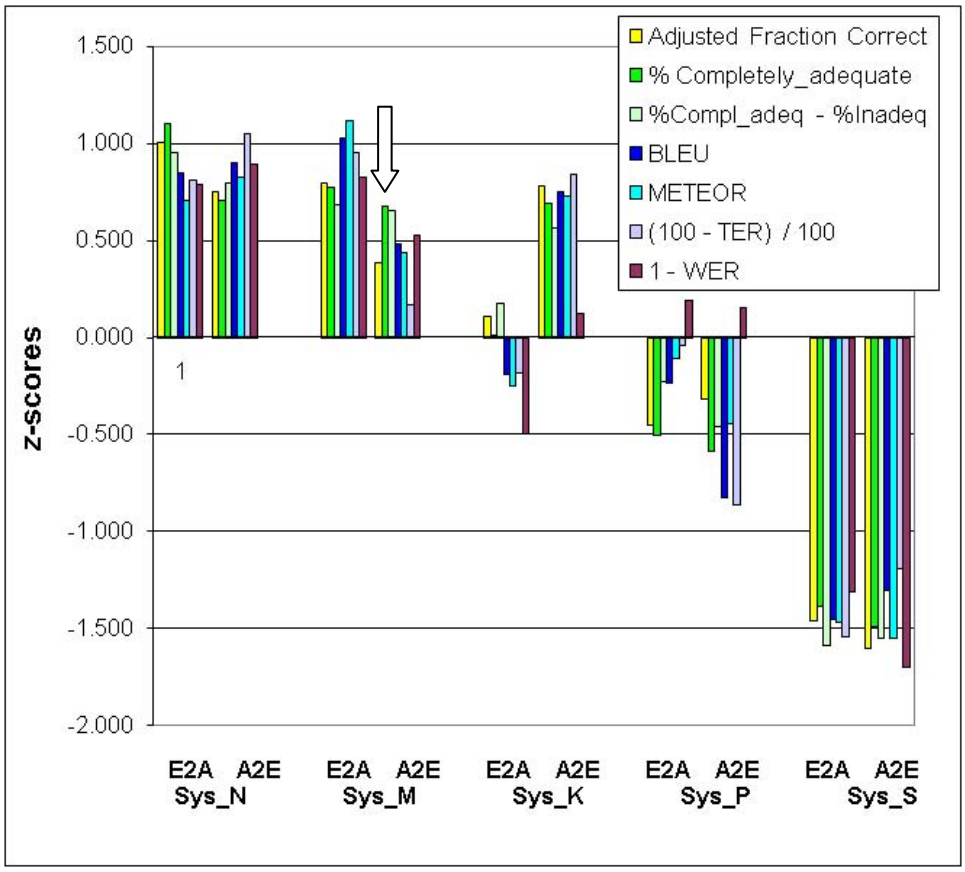

Figure 1: Synoptic overview of metrics for Arabic, converted to standard normal distribution *z*-scores.

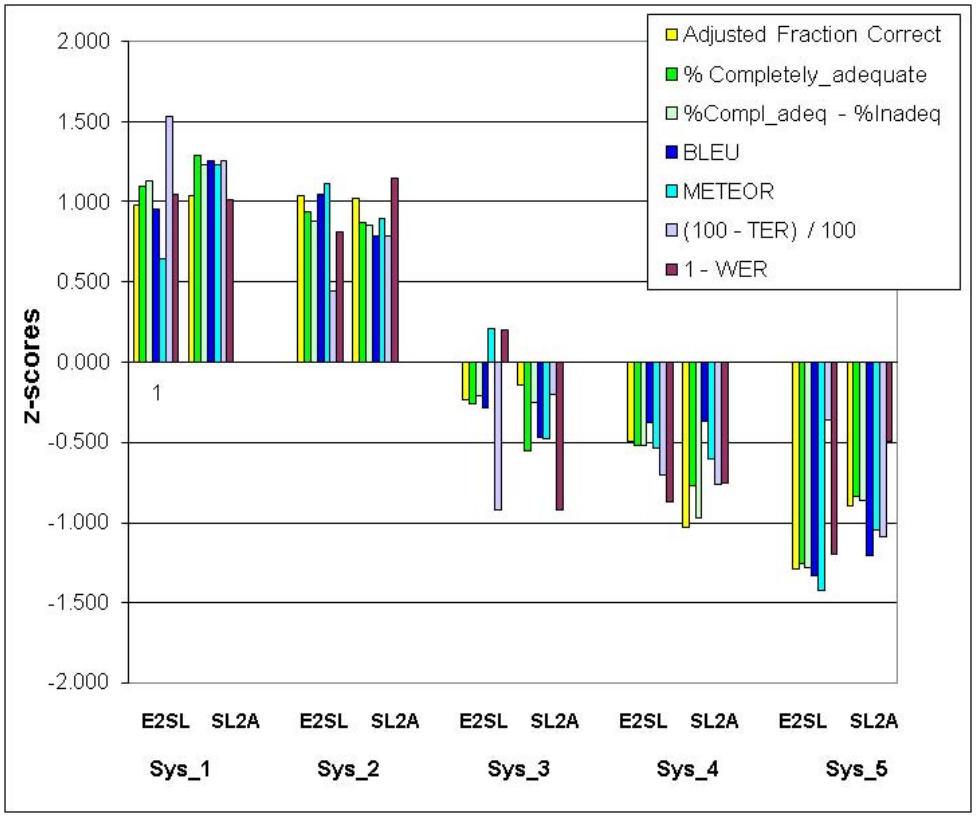

Figure 2: Synoptic overview of metrics for SurpriseLang, converted to standard normal distribution *z*-scores.

	BLEU	METEOR	1 - TER	LL_Odds	**AdjP. Corr.**	%Adeq	%Adeq - %Inad	1 - WER
BLEU	1							
METEOR	0.994	1						
1 - TER	0.994	0.993	1					
LL_Odds	0.955	0.919	0.928	1				
AdjP Corr.	0.972	0.944	0.959	0.982	1			
%Adeq	0.969	0.937	0.951	0.994	0.997	1		
%Adeq - %Inad	0.964	0.942	0.967	0.950	0.989	0.978	1	
1 - WER	0.958	0.968	0.974	0.872	0.887	0.888	0.900	1

Table 2. Pearson correlations for English to Iraqi Arabic

	BLEU	METEOR	1 - TER	LL_Odds	**AdjP Corr.**	%Adeq	%Adeq - %Inad	1 - WER
BLEU	1							
METEOR	0.974	1						
1 – TER	0.982	0.945	1					
LL_Odds	0.978	0.990	0.972	1				
AdjP Corr.	0.953	0.996	0.924	0.986	1			
%Adeq	0.979	0.988	0.930	0.971	0.978	1		
%Adeq - %Inad	0.966	0.991	0.917	0.965	0.982	0.994	1	
1 - WER	0.813	0.906	0.756	0.847	0.912	0.880	0.924	1

Table 3. Pearson correlations for Iraqi Arabic to English

	BLEU	METEOR	1 - TER	LL_Odds	**AdjP Corr.**	%Adeq	%Adeq - %Inad	1 - WER
BLEU	1							
METEOR	0.953	1						
1 – TER	0.730	0.542	1					
LL_Odds	0.983	0.921	0.810	1				
Adj. Corr.	0.998	0.958	0.746	0.989	1			
%Adeq	0.992	0.940	0.787	0.988	0.997	1		
%Adeq - %Inad	0.989	0.940	0.783	0.982	0.994	0.999	1	
1 – WER	0.930	0.949	0.704	0.932	0.951	0.956	0.961	1

Table 4. Pearson correlations for English to SurpriseLang

	BLEU	METEOR	1 - TER	LL_Odds	**AdjP Corr.**	%Adeq	%Adeq - %Inad	1 - WER
BLEU	1							
METEOR	0.988	1						
1 - TER	0.970	0.986	1					
LL_Odds	0.922	0.967	0.976	1				
Adj. Corr.	0.908	0.955	0.973	0.998	1			
%Adeq	0.961	0.990	0.976	0.974	0.960	1		
%Adeq - %Inad	0.933	0.972	0.987	0.992	0.990	0.983	1	
1 - WER	0.869	0.920	0.867	0.919	0.893	0.949	0.904	1

Table 5. Pearson correlations for SurpriseLang to English

9. Acknowledgments

We acknowledge Mari Maeda, the DARPA Program Manager for TRANSTAC, who has funded us to investigate these challenging, interesting scientific problems of how to evaluate multi-lingual speech-to-speech machine translation.

We wish to particularly acknowledge assistance from Jon Philips at MITRE who generated normalizations and ran the scoring software for Arabic and our surprise language, and we wish to acknowledge Arabic annotation assistance from Luma Ateyah of the Linguistic Data Consortium. We also acknowledge substantial helpful advice from Dan Parvaz of MITRE and Mohamed Maamouri of the Linguistic Data Consortium on various characteristics of Arabic. And of course we are grateful for the careful analyses from our panels of bilingual judges.

10. References

Banerjee, S., and A. Lavie (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL). Ann Arbor, MI.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, (10), pp. 37–46.

Condon, S., J. Phillips, C. Doran, J. Aberdeen, D. Parvaz, B. Oshika, G. Sanders, and C. Schlenoff (2008). Applying Automated Metrics to Speech Translation Dialogs. In *Proceedings of LREC-2008*.

Frederking, R. and S. Nirenburg (1994). Three Heads are Better Than One. In *Proceedings of the Fourth Conference on Applied Natural Language Processing (ANLP)*, pp. 95–100.

Frederking, R., A. W. Black, R. D. Brown, J. Moody, and E. Steinbrecher (2002). Field Testing the Tongues Speech-to-Speech Machine Translation System. In *Proceedings of LREC-2002* (Las Palmas, Gran Canaria, Canary Islands, Spain), pp. 160–164.

Hogan, C., and R. E. Frederking (1998). An Evaluation of the Multi-Engine MT Architecture. In *Proceedings of AMTA-1998*, pp. 113–124.

King, M. (1996). Evaluating Natural Language Processing Systems. In *Communications of the ACM* (39)1, pp. 73–79.

Knight, K. and I. Chander (1994). Automated Postediting of Documents. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Lavie, A., K. Sagae, and S. Jayaraman (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*. Washington, DC.

Lita, L. V., M. Rogati, and A. Lavie (2005). BLANC: Learning Evaluation Metrics for MT. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (*HLT/EMNLP*). (Vancouver, Canada, October 2005). pp. 740–747.

Melamed, I. D., R. Green, and J. P. Turian (2003). Precision and Recall of Machine Translation. *Proceedings of HLT/NAACL 2003* (Edmonton, Alberta, Canada), pp. 61–63.

Papineni, K., S. Roukos, T. Ward, and W. Zhu (2001). BLEU: A Method for Automatic Evaluation of Machine Translation. (IBM research report available at http://domino.watson.ibm.com/library/cybergig.nsf/).

Polzar, M., and E. Charniak (2006). Bllip: An Improved Evaluation Metric for Machine Translation. Brown University, Providence, Rhode Island. (Available as: http://www.cs.brown.edu/research/pubs/theses/masters/2006/mpozar.pdf)

Russo-Lassner, G., J. Lin, and P. Resnik (2005). A Paraphrase-based Approach to Machine Translation Evaluation. LAMP-TR-125 CS-TR-4754 UMIACS-TR-2005-57, University of Maryland, College Park, MD, August, 2005.

Snover M., B. Dorr, R. Schwartz, J. Makhoul, L. Micciulla, and R. Weischedel (2005). *A Study of Translation Error Rate with Targeted Human Annotation*. LAMP-TR-126, CS-TR-4755, UMIACS-TR-2005-58, University of Maryland, College Park, MD, July, 2005

Weiss, B., C. Schlenoff, G. Sanders, M. P. Steves, S. Condon, J. Phillips, and D. Parvaz (2008). Performance Evaluation of Speech Translation Systems. In *Proceedings of LREC-2008*.